Available online at [www.sciencedirect.com](http://www.sciencedirect.com/)

[Electronic Notes in Theoretical Computer Science 336 (2018) 135–153](https://doi.org/10.1016/j.entcs.2018.03.020)

[www.elsevier.com/locate/entcs](http://www.elsevier.com/locate/entcs)

A Categorical Characterization of Relative Entropy on Standard Borel Spaces

Nicolas Gagn´e[1](#_bookmark0)*,*[2](#_bookmark0) Prakash Panangaden[1](#_bookmark0)*,*[3](#_bookmark0)

*School of Computer Science McGill University Montr´eal, Qu´ebec, Canada*

**Abstract**

We give a categorical treatment, in the spirit of Baez and Fritz, of relative entropy for probability dis- tributions defined on standard Borel spaces. We define a category called **SbStat** suitable for reasoning about statistical inference on standard Borel spaces. We define relative entropy as a functor into Lawvere’s category [0*, ∞*] and we show convexity, lower semicontinuity and uniqueness.

*Keywords:* Entropy, G`ıry Monad, Bayesian Learning, standard Borel spaces

# Introduction

The inspiration for the present work comes from two recent developments. The first is the beginning of a categorical understanding of Bayesian inversion and learn- ing [[9,8,](#_bookmark33)[7](#_bookmark32)] the second is a categorical reconstruction of relative entropy [[3,](#_bookmark27)[2,](#_bookmark28)[15](#_bookmark40)]. The present paper provides a categorical treatment of entropy in the spirit of Baez and Fritz in the setting of standard Borel spaces, thus setting the stage to explore the role of entropy in learning.

Recently there have been some exciting developments that bring some categorical insights to probability theory and specifically to learning theory. These are reported in some recent papers by Clerc, Dahlqvist, Danos and Garnier [[9,](#_bookmark34)[8,](#_bookmark33)[7](#_bookmark32)]. The first of these papers showed how to view the Dirichlet distribution as a natural transforma- tion thus opening the way to an understanding of higher-order probabilities, while

1 This research has been supported by NSERC.

2 Email: [nicolas.gagne@mail.mcgill.ca](mailto:nicolas.gagne@mail.mcgill.ca)

3 Email: [prakash@cs.mcgill.ca](mailto:prakash@cs.mcgill.ca)

<https://doi.org/10.1016/j.entcs.2018.03.020>

1571-0661/© 2018 The Author(s). Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

the second gave a powerful framework for constructing several natural transforma- tions. In [[9](#_bookmark34)] the hope was expressed that one could use these ideas to understand Bayesian inversion, a core concept in machine learning. In [[7](#_bookmark32)] this was realized in a remarkably novel way. These papers carry out their investigations in the setting of standard Borel spaces and are based on the Giry monad [[11,](#_bookmark36)[13](#_bookmark38)].

In [[3,](#_bookmark27)[2](#_bookmark28)] a beautiful treatment of relative entropy is given in categorical terms. The basic idea is to understand entropy in terms of the results of experiments and observations. How much does one learn about a probabilistic situation by doing experiments and observing the results? A category is set up where the morphisms capture the interplay between the original space and the space of observations. In order to interpret the relative entropy as a functor they use Lawvere’s category which consists of a single object and a morphism for every extended positive real number [[14](#_bookmark39)].

Our contribution is to develop the theory of Baez et al. in the setting of standard Borel spaces; their work is carried out with finite sets. While the work of [[2](#_bookmark28)] gives a firm conceptual direction, it gives little guidance in the actual development of the mathematical theory. We had to redevelop the mathematical framework and find the right analogues for the concepts appropriate to the finite case.

# Background

In this section we review some of the background. We assume that the reader is familiar with concepts from topology and measure theory as well as basic category theory. We have found books by Ash [[1](#_bookmark26)], Billingsley [[4](#_bookmark29)] and Dudley [[10](#_bookmark35)] to be useful.

We will use letters like *X, Y, Z* for measurable spaces and capital Greek letters like Σ*,* Λ*,* Ω for *σ*-algebras. We will use *p, q, . . .* for probability measures. Given (*X,* Σ) and (*Y,* Λ) and a measurable function *f* : *X −→ Y* and a probability measure *p* on (*X,* Σ) we obtain a measure on (*Y,* Λ) by *p ◦ f−*1; this is called the *pushforward* measure or the *image* measure.

* 1. *The Giry monad*

We denote the category of measurable spaces and measurable functions by **Mes**. We recall the G`ıry [[11]](#_bookmark36) [4](#_bookmark1) functor Γ : **Mes** *→* **Mes** which maps each measurable space *X* to the space Γ(*X*) of probability measures over *X*. Let *A ∈* Σ, we define ev*A* : Γ(*X*) *−→* [0*,* 1] by ev*A*(*p*)= *p*(*A*). We endow Γ(*X*) with the smallest *σ*-algebra making all the ev’s measurable. A morphism *f* : *X −→ Y* in **Mes** is mapped to Γ(*f* ): Γ(*X*) *−→* Γ(*Y* ) by Γ(*f* )(*p*) = *p ◦ f−*1. With the following natural transformations, this endofunctor is a monad: the Giry monad. The natural transformation *η* : *I*

4 G`ıry’s name does have the accent grave on the ı; however, we omit it from now on to ease the typesetting.

*−→* Γ is given by *ηX* (*x*) = *δx*, the Dirac measure concentrated at *x*. The monad multiplication *μ* : Γ2 *→* Γ is given by

∫

*∀A ∈ B*(*X*)*, μX* (*p*)(*A*) := ev*A* d*p*

Γ(*X*)

where *p* is a probability measure in Γ(Γ(*X*)) and ev*A* : Γ(*X*) *−→* [0*,* 1] is the mea- surable function on Γ(*X*) defined by ev*A*(*p*)= *p*(*A*).

Even if **Mes** is an interesting category in and of itself, the need for regular condi- tional probabilities forces us to restrict ourselves to a subcategory of standard Borel spaces.

* 1. *Standard Borel spaces and disintegration*

The Radon-Nikodym theorem is the main tool used to show the existence of condi- tional probability distributions, also called Markov kernels, see the discussion below. It is a very general theorem, but it does not give as strong regularity features as one might want. A stronger theorem is needed; this is the so-called *disintegration theorem*. It requires stronger hypotheses on the space on which the kernels are being defined. A category of spaces that satisfy these stronger hypotheses is the category of standard Borel spaces. In order to define standard Borel spaces, we must first define Polish spaces.

**Definition 2.1** A *Polish space* is a separable, completely metrizable topological space.

**Definition 2.2** A *standard Borel space* is a measurable space obtained by forget- ting the topology of a Polish space but retaining its Borel algebra. The category of standard Borel spaces has measurable functions as morphisms; we denote it by **StBor**.

We can now state a version of the *disintegration theorem*. The following is also known as *Rohlin’s disintegration theorem* [[17](#_bookmark42)].

**Theorem 2.3 (Disintegration)** *Let* (*X, p*) *and* (*Y, q*) *be two standard Borel spaces equipped with probability measures, where q is the pushforward measure q* := *p ◦ f−*1 *for a Borel measurable function f* : *X → Y . Then, there exists a q-almost everywhere uniquely determined family of probability measures {py}y∈Y on X such that*

* + 1. *the function y '→ py*(*A*) *is a Borel-measurable function for each Borel- measurable set A ⊂ X;*
    2. *py is a probability measure on f−*1(*y*) *for q-almost all y ∈ Y ;*
    3. *for every Borel-measurable function h* : *X →* [0*, ∞*]*,*

∫ ∫ ∫

*h* d*p* = *h* d*py*d*q.*

*X Y f−*1(*y*)

The objects obtained are often called *regular conditional probability distributions*. One can find a crisp categorical formulation of disintegration in [[7](#_bookmark32), Theorem 1].

* 1. *The Kleisli category of* Γ *on* **StBor**

It is well known that the Giry monad on **Mes** restricted to **StBor** admits the same monad structure. [[11]](#_bookmark36)

The Kleisli category of Γ has as objects standard Borel spaces and as morphisms maps from *X* to Γ(*Y* ): *h* : *X −→* (*BY −→* [0*,* 1]) which are measurable. Here *BY* stands for the Borel sets of *Y* and Γ(*Y* ) has the *σ*-algebra described above. Now we can curry this to write it as *h* : *X × BY −→* [0*,* 1] or *h*(*x, U* ) where *x* is a point in *X* and *U* is a Borel set in *Y* . Written this way it is called a Markov kernel and one can view it as a transition probability function or conditional probability distribution given *x*. Composition of morphisms *f* : *X −→ Y* and *g* : *Y −→ Z* in the Kleisli category is given by the formula

∫

(*g ◦ f* )(*x, V ⊂ Z*)= *g*(*y, V* ) d*f* (*x, ·*)*.*

*Y*

For an arrow *s* : *Y →* Γ(*X*) in **StBor**, we write *sy* for *s*(*y*) or, in kernel form *s*(*y, ·*). For arrows *t* : *Z →* Γ(*Y* ) and *s* : *Y →* Γ(*X*) in **StBor**, we denote their Kleisli composition by *s* ˜*◦ t* := *μX ◦* Γ(*s*) *◦ t*. For standard Borel spaces equipped with a probability measure *p*, we sometimes omit the measure in the notation, *i.e.* we sometimes write *X* instead of (*X, p*). We say a probability measure *p* is *absolutely continuous* with respect to another measure *q* on the same measurable space *X*, denoted by *p q*, if for all measurable sets *B*, *q*(*B*) = 0 implies that *p*(*B*)= 0.

We note that absolute continuity is preserved by Kleisli composition; the proof is straightforward.

**Proposition 2.4** *Given a standard Borel space Y with probability measures q and qj such that q qj. Then, for arbitrary standard Borel space X and morphism s from Y to* Γ(*X*)*, we have s* ˜*◦ q s* ˜*◦ qj.*

# The categorical setting

In this section, following Baez and Fritz [[2](#_bookmark28)] (see also [[3](#_bookmark27)]) we describe the categories

**FinStat** and **FP** which they use for their characterization of entropy on finite

spaces. We then introduce the category **SbStat** which will be the arena for the generalization to standard Borel spaces.

Before doing so, we define the notion of coherence which will play an important role in what follows.

**Definition 3.1** Given standard Borel spaces *X* and *Y* with probability measure *p* and *q*, respectively, a pair (*f, s*), *f* : (*X, p*) *→* (*Y, q*) and *s* : *Y →* Γ(*X*), is said to be *coherent* when *f* is measure preserving, i.e., *q* = *p ◦ f−*1, and *sy* is a probability measure on *f−*1(*y*) *q*-almost everywhere. [5](#_bookmark4) If in addition, *p* is absolutely continuous with respect to *s* ˜*◦ q*, then we say that (*f, s*) is *absolutely coherent*.

**Definition 3.2** The category **FinStat** has

* **Objects** : Pairs (*X, p*) where *X* is a finite set and *p* a probability measure on *X*.
* **Morphisms** : Hom(*X, Y* ) are all coherent pairs (*f, s*), *f* : *X → Y* and *s* : *Y →*

Γ(*X*).

We compose arrows (*f, s*) : (*X, p*) *→* (*Y, q*) and (*g, t*) : (*Y, q*) *→* (*Z, m*) as follows: (*g, t*) *◦* (*f, s*) := (*g ◦ f, s* ˜*◦fin t*) where ˜*◦fin* is defined as

(*s* ˜*◦fin t*)*z*(*x*)= *tz*(*y*)*sy*(*x*)*.*

Σ

*y∈Y*

One can think of *f* as a measurement process from *X* to *Y* and of *s* as a hypothesis about *X* given an observation in *Y* . We say that a hypothesis *s* is *optimal* if *p* = *s* ˜*◦fin q*, or equivalenty, if *s* is a disintegration of *p* along *f* . We denote by **FP** the subcategory of **FinStat** consisting of the same objects, but with only those morphisms where the hypothesis is optimal. See [[3,2](#_bookmark28)] and [[15](#_bookmark40)] for a discussion of these ideas.

We now leave the finite world for a more general one: the category **SbStat**. **Definition 3.3** The category **SbStat** has

* **Objects** : Pairs (*X, p*) where *X* is a standard Borel space and *p* a probability measure on the Borel subsets of *X*.
* **Morphisms** : Hom(*X, Y* ) are all coherent pairs (*f, s*), *f* : *X → Y* and *s* : *Y →*

Γ(*X*).

We compose arrows (*f, s*) : (*X, p*) *→* (*Y, q*) and (*g, t*) : (*Y, q*) *→* (*Z, m*) as follows: (*g, t*) *◦* (*f, s*) := (*g ◦ f, s* ˜*◦ t*).

5 Note that (*f, s*) being coherent is equivalent to *ηY* = Γ(*f* ) *◦ s*.

Following the graphical representation from [[2](#_bookmark28)] we represent composition as follows:

*s t s* ˜*◦ t*

(*X, p*) (*Y, q*) (*Z, m*)

=======*⇒*

Composition

(*X, p*) (*Z, m*) *.*

*f g g ◦f*

**Proposition 3.4** *Given coherent pairs the composition is coherent. If, in addition, they are absolutely coherent, the composition is absolutely coherent.*

**Proof.** We first show that the composition is coherent, i.e., *ηZ* = (Γ(*g*) *◦* Γ(*f* )) *◦*

(*s* ˜*◦ t*). It is sufficient to show that the following diagram commutes:

Γ(*Y* ) *t Z*

Γ2(*X*)

*μX*

Γ2(*f* )

Γ(*ηY* )

Γ2(*Y* ) *ηZ*

Γ(*s*)

*μY*

Γ(*X*)

Γ(*f* )

Γ(*Y* )

Γ(*g*)

Γ(*Z*)

Using the hypothesis that *ηZ* = Γ(*g*) *◦ t* and the fact that Id = *μ ◦* Γ(*η*), we get that the right-hand square commutes. The triangle commutes since it is the application of Γ to our hypothesis *ηY* = Γ(*f* ) *◦ s* and the left-hand square commutes because *μ* is a natural transformation. Therefore, the whole diagram commutes and we have thus shown the composition of coherent morphisms is also coherent.

Next, in addition, assume the pairs (*f, s*) and (*g, t*) are absolutely coherent. We show *p* (*s* ˜*◦ t* ˜*◦ m*). By hypotheses, *p s* ˜*◦ q* and *q t* ˜*◦ m*. Using Proposition

[2.4](#_bookmark3) on *q t* ˜*◦ m*, we get *s* ˜*◦ q s* ˜*◦ t* ˜*◦ m*. By transitivity of, we conclude

*p* (*s* ˜*◦ t* ˜*◦ m*). *2*

We end this section by defining one more category; this one is due to Lawvere [[14](#_bookmark39)]. It is just the set [0*, ∞*] but endowed with categorical structure. This allows numerical values associated with morphisms to be regarded as functors.

**Definition 3.5** The category [0*, ∞*] has

* **Objects** : One single object: *•*.
* **Morphisms** : For each element *r ∈* [0*, ∞*], one arrow *r* : *•→ •*. Arrow composition is defined as addition in [0*, ∞*].

This is a remarkable category with monoidal closed structure and many other in- teresting properties.

# Relative entropy functor

We recapitulate the definition of the relative entropy functor on **FinStat** from Baez and Fritz [[2](#_bookmark28)] and then extend it to **SbStat**.

**Definition 4.1** The relative entropy functor *REfin* is defined from **FinStat** to [0*, ∞*] as follows:

* + **On Objects** : It maps every object (*X, p*) to *•*.
  + **On Morphisms** : It maps a morphism (*f, s*): (*X, p*) *→* (*Y, q*) to *Sfin*(*p, s* ˜*◦fin q*), where

Σ

*Sfin*

(*p, s* ˜*◦fin*

1. := *p*(*x*) ln *p*(*x*) *.*

*x∈X* (*s* ˜*◦fin q*)(*x*)

The convention from now on will be that *∞· c* = *c ·∞* = *∞* for 0 *< c ≤ ∞* and

*∞·* 0=0 *·∞* = 0. We extend *REfin* from **FinStat** to **SbStat**.

**Definition 4.2** The relative entropy functor *RE* is defined from **SbStat** to [0*, ∞*] as follows:

* + **On Objects** : It maps every object (*X, p*) to *•*.
  + **On Morphisms** : Given a coherent morphism (*f, s*): (*X, p*) *→* (*Y, q*), if (*f, s*) is absolutely coherent, then *RE*((*f, s*)) = *S*(*p, s* ˜*◦ q*) , where

*S*(*p, s* ˜*◦ q*) :=

∫

*X*

log d*p* d*p,* d(*s* ˜*◦ q*)

otherwise it is defined as *RE*((*f, s*)) = *∞*.

This quantity is also known as the *Kullback-Leibler divergence*.

We could have defined our category to have only absolutely coherent morphisms but it would make the comparison with the finite case more awkward as the finite case does not assume the morphisms to be absolutely coherent. The present definition leads to slightly awkward proofs where we have to consider absolutely coherent pairs and ordinary coherent pairs separately.

Clearly, *RE* restricts to *REfin* on **FinStat**. If (*f, s*) is absolutely coherent, then *p* is absolutely continuous with respect to (*s* ˜*◦ q*) and the Radon-Nikodym derivative is defined. The relative entropy is always non-negative [[12](#_bookmark37)]; this is an easy consequence of Jensen’s inequality. This shows that RE is defined everywhere in **SbStat**.

We will use the following notation occasionally:

*s*

*RE* (*X, p*)  (*Y, q*) := *RE*((*f, s*))*.*

*f*

It remains to show that *RE* is indeed a functor. That is, we want to show that

*s*

*RE* (*X, p*)

*f*

(*Y, q*)

*t*

*s*

(*Z, m*) = *RE* (*X, p*)

*g f*

(*Y, q*) + *RE* (*Y, q*)

*t*

(*Z, m*) *.*

*g*

In order to do so, we will need the following lemma.

**Lemma 4.3** *The relative entropy is preserved under pre-composition by optimal hypotheses, i.e., for any* (*g, t*): (*Y, q*) *→* (*Z, m*) *and* (*f, s*): (*X, s* ˜*◦ q*) *→* (*Y, q*)*, we have*

*t s t*

*RE* (*Y, q*) (*Z, m*) = *RE* (*X, s* ˜*◦ q*) (*Y, q*) (*Z, m*) *.*

*g f g*

## Proof.

**Case I :** (*g, t*) **is absolutely coherent.** Since (*g, t*) is absolutely coherent, so is (*g ◦ f, s* ˜*◦ t*) by Proposition [2.4](#_bookmark3). Hence, to show *RE*(*g, t*) = *RE*(*g ◦ f, s* ˜*◦ t*) is to show

∫ log d*q*  d*q* = ∫ log d(*s* ˜*◦ q*) d(*s* ˜*◦ q*)*.*

*Y*

*X*

d(*t* ˜*◦ m*)

d(*s* ˜*◦ t* ˜*◦ m*)

Because *f* is measure preserving, it is sufficient to show that the following functions on *X*

d*q* d(*t* ˜*◦ m*)

* *f* =

d(*s* ˜*◦ q*) d(*s* ˜*◦ t* ˜*◦ m*)

*s* ˜*◦ t* ˜*◦ m*-almost everywhere.

By the Radon-Nikodym theorem, it is sufficient to show that for any *E ⊂ X* mea- surable set, we have

(*s* ˜*◦ q*)(*E*)= ∫

d*q*

*◦ f* d(*s* ˜*◦ t* ˜*◦ m*)*.*

*E* d(*t* ˜*◦ m*)

The following calculation establishes the above.

d*q*

∫

*◦ f* d(*s* ˜*◦ t* ˜*◦ m*) (1)

*E* d(*t* ˜*◦ m*)

= ∫

∫

*Y*

∫

*x∈f*

*−*1(*y*)*∩E*

d*q* d(*t* ˜*◦ m*)

* *f* (*x*) d(*s* ˜*◦ t* ˜*◦ m*)*y* d(*t* ˜*◦ m*) (2)

d*q*

=

*Y* d(*t* ˜*◦ m*)

∫

d*q*

=

*Y* d(*t* ˜*◦ m*)

(*y*) ∫

(*y*) ∫

*f*

*f*

*−*1(*y*)*∩E*

*−*1(*y*)*∩E*

d(*s* ˜*◦ t* ˜*◦ m*)*y* d(*t* ˜*◦ m*) (3)

d*sy* d(*t* ˜*◦ m*) (4)

= ∫ d*q* (*y*)*s* (*E ∩ f−*1(*y*)) d(*t* ˜*◦ m*) (5)

*Y*

d(*t* ˜*◦ m*)

*y*

∫

d*q*

= (*y*)*s* (*E*) d(*t* ˜*◦ m*) (6)

*Y*

d(*t* ˜*◦ m*)

*y*

= *sy*(*E*) d*q* (7)

*Y*

∫

= (*s* ˜*◦ q*)(*E*) (8)

We get ([2](#_bookmark7)) by applying the disintegration theorem to *f* : (*X, s* ˜*◦ t* ˜*◦ m*) *→* (*Y, t* ˜*◦ m*). The equation ([3](#_bookmark8)) follows by using the fact that d*q ◦ f* is constant on *f−*1(*y*) for every *y*. To obtain ([4](#_bookmark9)) we apply Lemma [A.1](#_bookmark43). To show ([6](#_bookmark10)) we use the fact that *sy* is a probability measure on *f−*1(*y*). We get ([7](#_bookmark11)) by the definition of the Radon-Nikodym derivative and we finally establish ([8](#_bookmark12)) by the definition of Kleisli composition.

d(*t* ˜*◦ m*)

**Case II :** (*g, t*) **is not absolutely coherent.** We have *RE*((*g, t*)) = *∞*. We show that (*g ◦ f, s* ˜*◦ t*) is not absolutely coherent, i.e., *s* ˜*◦ q* is not absolutely continuous with respect to *s* ˜*◦ t* ˜*◦ m*.

Since, by hypothesis, *q t* ˜*◦ m* doesn’t hold, there exists a measurable set *B ⊂ Y* such that (*t* ˜*◦ m*)(*B*) = 0 but *q*(*B*) *>* 0. We argue that (*s* ˜*◦ t* ˜*◦ m*)(*f−*1(*B*)) = 0 and (*s* ˜*◦ q*)(*f−*1(*B*)) *>* 0. On one hand, we have

∫

(*s* ˜*◦ t* ˜*◦ m*)(*f−*1(*B*)) = *sy*(*f−*1(*B*))d(*t* ˜*◦ m*) *≤* (*t* ˜*◦ m*)(*B*)= 0*.*

*B*

But on the other hand, since *f* is a measure preserving map from (*X, s* ˜*◦ q*) to (*Y, q*), we have (*s* ˜*◦ q*)(*f−*1(*B*)) = *q*(*B*) *>* 0.

Therefore,

*RE*((*g, t*)) = *∞* = *RE*((*g ◦ f, s* ˜*◦ t*))*.* *2*

**Theorem 4.4 (Functoriality)** *Given arrows* (*f, s*) : (*X, p*) *→* (*Y, q*) *and* (*g, t*) : (*Y, q*) *→* (*Z, m*)*, we have*

*RE* ((*g, t*) *◦* (*f, s*)) = *RE*((*f, s*)) + *RE*((*g, t*))*.*

**Proof.** Note that by definition, *RE* ((*g, t*) *◦* (*f, s*)) = *RE* ((*g ◦ f, s* ˜*◦ t*)).

**Case I :** (*f, s*) **and** (*g, t*) **are absolutely coherent.** By Proposition [3.4](#_bookmark5), we have that (*g ◦ f, s* ˜*◦ t*) is absolutely coherent.

*RE* ((*g ◦ f, s* ˜*◦ t*)) =

∫

*X*

∫

log d*p* d*p* (9)

d(*s* ˜*◦ t* ˜*◦ m*)

= log d*p* d(*s* ˜*◦ q*) d*p* (10)

*X* d(*s* ˜*◦ q*) d(*s* ˜*◦ t* ˜*◦ m*)

= ∫ log d*p*  d*p* + ∫ log d(*s* ˜*◦ q*) d*p* (11)

∫

d(*s* ˜*◦ q*)

d(*s* ˜*◦ t* ˜*◦ m*)

*X*

*X*

= *RE*((*f, s*)) +

*X*

log d(*s* ˜*◦ q*) d*p* (12) d(*s* ˜*◦ t* ˜*◦ m*)

= *RE*((*f, s*)) + *RE*((*g, t*)) (13)

We get ([10](#_bookmark14)) by the chain rule for Radon-Nikodym derivatives and ([13](#_bookmark15)) by applying Lemma [4.3](#_bookmark6).

**Case II :** (*g, t*) **is not absolutely coherent.** We argue that (*g ◦ f, s* ˜*◦ t*) is not absolutely coherent. By hypothesis, *q t* ˜*◦ m* doesn’t hold, so there is a measurable set *B ⊂ Y* such that (*t* ˜*◦ m*)(*B*) = 0 and *q*(*B*) *>* 0. We show that (*s* ˜*◦ t* ˜*◦ m*) *f−*1(*B*) = 0 and *p*(*f−*1(*B*)) *>* 0. On one hand, we have

(*s* ˜*◦ t* ˜*◦ m*) *f−*1(*B*) = ∫

*sy* *f−*1(*B*) d(*t* ˜*◦ m*) *≤* (*t* ˜*◦ m*)(*B*)= 0*,*

*B*

but on the other hand, we have *p*(*f−*1(*B*)) = *q*(*B*) *>* 0. Therefore

*RE* ((*g, t*) *◦* (*f, s*)) = *∞* = *RE*((*f, s*)) + *RE*((*g, t*))*.*

## Case III : (*f, s*) is not absolutely coherent.

Although we relegated the proof of case III to the appendix, it is neither trivial nor boring. [6](#_bookmark16)

For both of the above cases, we can deduce that

*RE* ((*g, t*) *◦* (*f, s*)) = *∞* = *RE*((*f, s*)) + *RE*((*g, t*))*.*

*2*

6 It is not analogous to the previous case since the existence of a measurable set *A ⊂ X* such that (*s* ˜*◦ q*)(*A*) = 0 and *p*(*A*) *>* 0 is surprisingly not enough to conclude that (*s* ˜*◦ t* ˜*◦ m*)(*A*)= 0.

We have thus shown that *RE* is a well-defined functor from **SbStat** to [0*, ∞*].

* 1. *Convex linearity*

We show below that the relative entropy functor satisfies a convex linearity property. In [[2](#_bookmark28)] convexity looks familiar; here since we are performing “large” sums we have to express it as an integral. First we define a localized version of the relative entropy.

Note that Lemma [A.1](#_bookmark43) in the appendix says that *sy* = (*s* ˜*◦ q*)*y q*-almost everywhere. Thus, in the following there is no notational clash between the kernel *sy* and (*s* ˜*◦ q*)*y*, the later being the disintegration of (*s* ˜*◦ q*) along *f* .

Given an arrow (*f, s*): (*X, p*) *→* (*Y, q*) in **StBor** and a point *y ∈ Y* , we denote by (*f, s*)*y*, the morphism (*f, s*) restricted to the pair of standard Borel spaces *f−*1(*y*) and *{y}*. Explicitly,

(*f, s*)*y* := ( *f| −*1 *, sy*): (*f−*1(*y*)*, py*) *−→* (*{y}, δy*)*,*

*f* (*y*)

where *δy* is the one and only probability measure on *{y}*.

**Definition 4.5** A functor *F* from **SbStat** to [0*, ∞*] is *convex linear* if for every arrow (*f, s*): (*X, p*) *→* (*Y, q*), we have

*F* ((*f, s*)) = *F* ((*f, s*)*y*) d*q.*

∫

*Y*

We will sometimes refer to the relative entropy of (*f, s*)*y* as the *local relative entropy of* (*f, s*) *at y*.

**Theorem 4.6 (Convex Linearity)** *The functor RE is convex linear, i.e., for ev- ery arrow* (*f, s*): (*X, p*) *→* (*Y, q*)*, we have*

*RE*((*f, s*)) = *RE* ((*f, s*)*y*) d*q.*

∫

*Y*

## Proof.

**Case I :** (*f, s*) **is absolutely coherent.**

We have

*RE*((*f, s*)) =

∫

*X*

log d*p* d*p* (14)

d(*s* ˜*◦ q*)

= ∫ ∫

log d*p*  d*p*

d*q* (15)

*−*1

= ∫ ∫ log d*py*  d*p* d*q* (16)

*Y*

*f*

(*y*)

d(*s* ˜*◦ q*)

*y*

*Y*

*f*

*−*1

(*y*)

d(*s* ˜*◦ q*)*y*

*y*

∫

= *RE*((*f, s*)*y*) d*q.* (17)

*Y*

We get ([15](#_bookmark17)) by the disintegration theorem and ([16](#_bookmark18)) by applying Lemma [A.2](#_bookmark44).

**Case II :** (*f, s*) **is not absolutely coherent.** By the hypothesis of (*f, s*) not being absolutely coherent, there is a measurable set *A ⊂ X* such that (*s* ˜*◦ q*)(*A*)=0 and *p*(*A*) *>* 0. Applying lemma [A.1,](#_bookmark43) on one hand we have

∫ ∫

(*s* ˜*◦ q*)*y*(*A*) d*q* = *sy*(*A*) d*q* = (*s* ˜*◦ q*)(*A*)= 0*,*

*Y Y*

but on the other hand we have

∫

*py*(*A*) d*q* = *p*(*A*) *>* 0*.*

*Y*

Hence, the subset of *Y* on which *py* (*s* ˜*◦ q*)*y* doesn’t hold contains a set of measure strictly greater than 0. Therefore,

∫

*RE*((*f, s*)) = *∞* = *RE* ((*f, s*)*y*) d*q.* *2*

*Y*

* 1. *Lower-semi-continuity*

Recall that a sequence of probability measures *pn* converges strongly to *p*, denoted by *pn → p*, if for all measurable set *E*, one has lim*n→∞ pn*(*E*)= *p*(*E*).

**Definition 4.7** A functor *F* from **SbStat** to [0*, ∞*] is *lower semi-continuous* if for every arrow (*f, s*): (*X, p*) *→* (*y, δy*), whenever *pn → p* and *sn → s*, then

*s sn*

*F* (*X, p*)  (*y, δy*) *≤* lim inf *F* (*X, pn*) (*y, δy*) *.*

*n→∞*

*f f*

Note that a lower semi-continuous functor *F* on **SbStat** restricts to a lower semi- continuous functor [7](#_bookmark19) on **FinStat**.

7 As defined slightly differently in [[2](#_bookmark28)].

**Theorem 4.8 (Lower semi-continuity)** *The functor RE is lower semi- continuous.*

**Proof.** This is a direct consequence of Pinsker [[16](#_bookmark41), Section 2.4]. *2*

# Uniqueness

We now show that the relative entropy is, up to a multiplicative constant, the unique functor satisfying the conditions established so far. We first prove a crucial lemma.

**Lemma 5.1** *Let X be a Borel space equipped with probability measures p and q, if*

*p q, then we can ﬁnd a sequence of simple functions p∗ on X such that for the*

*n*

*sequence of probability measures pn*(*E*) := ∫*E p* d*q, we have that pn and p agree on*

*∗*

*n*

*the elements of the partition on X induced by p∗ and moreover, pn → p strongly.*

*n*

**Proof.** We write *In,k* for the interval [*k*2*−n,* (*k* + 1)2*−n*) and *In,≤* for the interval [*n, ∞*). Denote by *Kn* the index set *{*0*,* 1*,..., n*2*n −* 1*, ≤}* of *k*. We fix a version

d*p* of the Radon-Nikodym such that d*p*

d*q*

d*q*

*< ∞* everywhere. We define a family of

partitions and a family of simple functions as follows:

*X* := *xj ∈ X |* d*p* (*xj*) *∈ I*  *, p∗* (*x*) := *p* (*Xn,k*) on *x ∈ X .*

*n,k*

d*q*

*n,k*

*n*

*q* (*Xn,k*)

*n,k*

Every function induces a partition on the domain; if moreover the function is simple, the induced partition is finite.

We first note that *pn* and *p* agree on the elements of the partition induced by *p∗* :

*n*

*n*

*p* (*X*

*n*

*n,k*

)= ∫

*p∗* d*q* = ∫

*p* (*Xn,k*) d*q* = *p* (*Xn,k*) *q*(*X*

)= *p*(*X* )*.*

*n,k*

*n,k*

*Xn,k*

*Xn,k*

Next, we prove the strong convergence of *pn → p*. We first show *p∗ →* d*p* pointwise.

*q* (*Xn,k*)

*q* (*Xn,k*)

*n* d*q*

Let *x ∈ X*. Pick *N* large enough such that d*p* (*x*) *≤ N* . For a fixed integer *n ≥ N* ,

d*q*

there is exactly one *kn* for which *x ∈ Xn,kn*

. On the one hand, we have *kn*2*−n ≤*

d*p* (*x*) *≤* (*kn* + 1)2*−n* on *Xn,k* . But on the other hand, by integrating over *Xn,k*

d*q n n*

and dividing everything by *q*(*X*

*n,kn*

), we also have *kn*

2*−n ≤ p*(*Xn,kn* ) *≤* (*k*

*q*(*Xn,kn* )

*n*

+ 1)2*−n*

on *Xn,kn* . We thus get pointwise convergence since we have

*p* (*x*) *−* (*x*) = *−* (*x*) *≤* 2 for any *n ≥ N.*

*∗* d*p p* (*Xn,kn* ) d*p −n*

*n*

d*q*

*q* (*Xn,kn*

)

d*q*

From the above inequality and the choice of *N* , we note the following

*p∗* (*x*) *≤* d*p* (*x*)+ 2*−n ≤* d*p* (*x*)+ 1*,* for *x* with d*p* (*x*) *< n,*

*n* d*q* d*q* d*q*

*p∗* (*x*)= *p*(*Xn,≤*) *≤* 1 *≤* d*p* (*x*)+ 1*,* for *x* with d*p* (*x*) *≥ n.*

*n* d*q* d*q*

So for all *n*, we can bound *p∗* (*x*) everywhere by the integrable function *g*(*x*) := d*p* (*x*) + 1. Given a measurable set *E ⊂ X*, we can thus apply Lebesgue’s dominated convergence theorem. We get

*n*

d*q*

lim

*n→∞*

*pn*(*E*) = lim

*E*

*n→∞* ∫

*∗* lim

*E n→∞*

*p* d*q* = ∫

*n*

*∗* d*p*

*n* d*q* = *p*(*E*)*.*

*p* d*q* = ∫

*E* d*q*

*2*

Before proving uniqueness, we recall the main theorem of Baez and Fritz [[2](#_bookmark28)] on

## FinStat.

**Theorem 5.2** *Suppose that a functor*

*F* : **FinStat** *→* [0*, ∞*]

*is lower semicontinuous, convex linear and vanishes on* **FP***. Then for some* 0 *≤*

*c ≤∞ we have F* (*f, s*)= *cREfin*(*f, s*) *for all morphisms* (*f, s*) *in* **FinStat***.*

We are now ready to extend this characterization to **SbStat**. **Theorem 5.3** *Suppose that a functor*

*F* : **SbStat** *→* [0*, ∞*]

*is lower semicontinuous, convex linear and vanishes on* **FP***. Then for some* 0 *≤*

*c ≤∞ we have F* (*f, s*)= *cRE*(*f, s*) *for all morphisms.*

**Proof.** Since *F* satisfies all the above properties on **FinStat**, we can apply The- orem [5.2](#_bookmark21) in order to establish that *F* = *cREfin* = *cRE* for all morphisms in the subcategory **FinStat**. We show that *F* extends uniquely to *cRE* on all morphisms in **SbStat**.

By convex linearity of *F* , for an arbitrary morphism (*f, s*) from (*X, p*) to (*Y, q*), we have

∫

*F* ((*f, s*)) =

*Y*

*F* ((*f, s*)*y*) d*q,*

so *F* is totally described by its local relative entropies. It is thus sufficient to show

*F* = *cRE* on an arbitrary morphism (*f, s*) : (*X, p*) *→* (*{y}, δy*). The case where *p* is not absolutely continuous with respect to *s* is straightforward, so let us assume *p s*. [8](#_bookmark22)

8 See the extended version for details.

We apply Lemma [5.1](#_bookmark20) with *p* and *s* to get the family of simple functions *p∗* and the corresponding family of partitions *{Xn,k}*. We define *πn* as the function that maps *x ∈ Xn,k′* to the element *Xn,k′ ∈ {Xn,k}k∈Kn* . Denote by *sπn* the disintegration of *s* along *πn* and by *sn* the corresponding marginal. Note that since *pn* and *p* agree on every *Xn,k*, *pn* is indeed the push-forward of *p* along *πn*, so we can identify *pn* to the corresponding marginal of *p*. Presented as diagrams, we have

*n*

*sπn sn s*

(*X, p*) (*{Xn,k}, pn*) (*{y}, δy*)

=======*⇒*

Composition

(*X, p*) (*{y}, δy*)*.*

*πn fn f*

From the above diagram and the hypothesis that *F* is a functor, we have the fol- lowing inequality

*F* ((*fn, sn*)) *≤ F* ((*f, s*)), for all *n ∈* N*.* (18) Note that, on the one hand the disintegration of *pn* along *πn* at the point

*Xn,k′ ∈ {Xn,k}* is given by *pn,π* := *pn*(*·*)*/pn*(*Xn,k′* ), but on the other hand, for any measurable set *E ⊂ X*, we also have

*k*Σ*∈Kn*

*E*

∫

*Xn,k*

d*pn,π*

*sn*

(*Xn,k*

)=

*k∈Kn*

Σ

*pn*(*E ∩ Xn,k*) *s pn*(*Xn,k*) *n*

(*Xn,k*)

=

Σ

*k∈Kn*

*s*(*E ∩ Xn,k*) *s s*(*Xn,k*) *n*

(*Xn,k*

)=

*k∈Kn*

Σ

*s*(*E ∩ X*

*n,k*

)= *s*(*E*)*.*

This means that *pn,π* is the disintegration of *s* along *πn*. Presented as diagrams, where we use *fpn* instead of *f* to indicate that the arrow leaves from the object (*X, pn*) as opposed to (*X, p*), we have

*pn,π sn s*

(*X, pn*) (*{Xn,k}, pn*) (*{y}, δy*)

=======*⇒*

Composition

(*X, pn*) (*{y}, δy*)*.*

*πn fn*

*fpn*

But since *F* vanishes on **FP**, we have *F* ((*πn, pn,π*)) = 0. Combined with the fact that *F* is a functor, we get

*F* ((*fpn , s*)) = *F* ((*πn, pn,π*)) + *F* ((*fn, sn*)) = *F* ((*fn, sn*)) *.* (19)

By Lemma [5.1](#_bookmark20), we know that *pn → p*, in terms of our diagrams we have

*s*

(*X, pn*) (*{y}, δy*)

*fpn*

============*⇒*

Strong Convergence

*s*

(*X, p*) (*{y}, δy*)*.*

*f*

Hence, combining ([19](#_bookmark24)) with the lower semicontinuity of *F* , we also have the inequal- ity

*F* ((*f, s*)) *≤* lim inf *F* ((*fpn , s*)) = lim inf *F* ((*fn, sn*))*.* (20)

*n→∞ n→∞*

Since (*fn, sn*) is in **FinStat**, we must have *F* ((*fn, sn*)) = *cRE*((*fn, sn*)). Thus, combining ([18](#_bookmark23)) and ([20](#_bookmark25)), we get that *F* ((*f, s*)) must satisfy

lim sup *cRE*((*fn, sn*)) *≤ F* ((*f, s*)) *≤* lim inf *cRE*((*fn, sn*))*,*

*n→∞*

but so does *cRE*((*f, s*)). We also have

*n→∞*

lim sup *cRE*((*fn, sn*)) *≤ cRE*((*f, s*)) *≤* lim inf *cRE*((*fn, sn*))*.*

*n→∞*

*n→∞*

Therefore *F* ((*f, s*)) = *cRE*((*f, s*)), as desired. *2*

# Conclusions and Further Directions

As promised, we have given a categorial characterization of relative entropy on standard Borel spaces. This greatly broadens the scope of the original work by Baez et al. [[3,](#_bookmark27)[2](#_bookmark28)]. However, the main motivation is to study the role of entropy arguments in machine learning. These appear in various ad-hoc ways in machine learning but with the appearance of the recent work by Danos and his co-workers [[9,](#_bookmark34)[7,](#_bookmark32)[8](#_bookmark33)] we feel that we have the prospect of a mathematically well-defined framework on which to understand Bayesian inversion and its interplay with entropy. The most recent paper in this series [[7](#_bookmark32)] adopts a point-free approach introduced in [[5,](#_bookmark30)[6](#_bookmark31)]. It would be interesting to extend our definitions to a point-free situation.

# Acknowledgments

We have benefitted from discussions with Florence Clerc, Vincent Danos, Tobias Fritz, Renaud Raqu´epas and Ilias Garnier. We are grateful to the anonymous referees for very helpful comments. This research was supported by a grant from Google and from NSERC.

# References

1. Ash, R. B., “Real Analysis and Probability,” Academic Press, 1972.
2. Baez, J. C. and T. Fritz, *A bayesian characterization of relative entropy*, Theory and Applications of Categories **29** (2014), pp. 422–456.
3. Baez, J. C., T. Fritz and T. Leinster, *A characterization of entropy in terms of information loss*, Entropy **13** (2011), pp. 1945–1957.
4. Billingsley, P., “Probability and Measure,” Wiley-Interscience, 1995.
5. Chaput, P., V. Danos, P. Panangaden and G. Plotkin, *Approximating Markov processes by averaging*, in: *Proceedings of the 37th International Colloquium On Automata Languages And Programming (ICALP)*, Lecture Notes In Computer Science **5556**, 2009, pp. 127–138.
6. Chaput, P., V. Danos, P. Panangaden and G. Plotkin, *Approximating Markov processes by averaging*,

J. ACM **61** (2014), pp. 5:1–5:45.

URL <http://doi.acm.org/10.1145/2537948>

1. Clerc, F., V. Danos, F. Dahlqvist and I. Garnier, *Pointless learning*, in: *Proceedings of FoSSaCS 2017*, 2017.
2. Dahlqvist, F., V. Danos and I. Garnier, *Giry and the machine*, Electronic Notes in Theoretical Computer Science **325** (2016), pp. 85–110.
3. Danos, V. and I. Garnier, *Dirichlet is natural*, Electronic Notes in Theoretical Computer Science **319**

(2015), pp. 137–164.

1. Dudley, R. M., “Real Analysis and Probability,” Wadsworth and Brookes/Cole, 1989.
2. Giry, M., *A categorical approach to probability theory*, in: B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, number 915 in Lecture Notes In Mathematics (1981), pp. 68–85.
3. Kullback, S. and R. A. Leibler, *On information and sufficiency*, The annals of mathematical statistics

**22** (1951), pp. 79–86.

1. Lawvere, F. W., *The category of probabilistic mappings* (1964), unpublished typescript.
2. Lawvere, F. W., *Metric spaces, generalized logic and closed categories*, Rend. Sem. Mat. Fis. Milano

**43** (1973), pp. 135–166.

1. Leinster, T., *An operadic introduction to entropy*, n-category cafe.
2. Pinsker, M. S., *Information and information stability of random variables and processes* (1960).
3. Rokhlin, V. A., *On the fundamental ideas of measure theory*, Matematicheskii Sbornik **67** (1949),

pp. 107–150.

# A Lemmas and Proofs

**Lemma A.1** *Given an arrow* (*f, s*): (*X, p*) *→* (*Y, q*) *in* **SbStat***. Let {*(*s* ˜*◦ q*)*y}y∈Y*

*be a disintegration of* (*s* ˜*◦ q*) *along f, then*

(*s* ˜*◦ q*)*y* = *sy q-almost everywhere.*

We just have to show that *{sy}y∈Y* satisfies the three properties implied by the disintegration theorem. We prove the third one; the first two being obvious.

[2.3](#_bookmark2) (iii) *: For every Borel-measurable function h* : *X →* [0*, ∞*]*,*

∫*X h* d(*s* ˜*◦ q*)= ∫*y∈Y* ∫*f−*1(*y*) *h* d*sy* d*q.*

**Proof.** Let’s assume as a special case that *h* is the indicator function for a mea- surable set *E ⊂ X*. Then, we have

∫ ∫ ∫ ∫ ∫

*h* d(*s* ˜*◦ q*)= d(*s* ˜*◦ q*)= (*s* ˜*◦ q*)(*E*)= *sy*(*E*) d*q* = *h* d*sy* d*q.*

*X E y∈Y y∈Y ƒ−*1(*y*)

We have shown that it is true for any indicator function. By linearity, it is true for any simple function and then, by the monotone convergence theorem, it is true for all Borel-measurable functions *h* : *X →* [0*, ∞*]. *2*

## Proof of Case III of Theorem [4.4](#_bookmark13)

By the hypothesis of (*f, s*) not being absolutely coherent, *p s* ˜*◦ q* doesn’t hold, so there is a measurable set *A ⊂ X* such that (*s* ˜*◦ q*)(*A*)=0 and *p*(*A*) *>* 0.

We partition *A* into

*Aє* := *{x ∈ A | sƒ* (*x*)(*A*) *>* 0*}* and *A*0 := *{x ∈ A | sƒ* (*x*)(*A*)= 0*}* and we partition *Y* into

*Bє* := *{y ∈ Y | sy*(*A*) *>* 0*}* and *B*0 := *{y ∈ Y | sy*(*A*)= 0*}.*

We argue that (*s* ˜*◦ t* ˜*◦ m*)(*A*0)=0 and *p*(*A*0) *>* 0.

Since *A*0 *⊂ f−*1(*B*0), *f−*1(*Bє*) is disjoint from *A*0, so for all *y ∈ Bє* we have *sy*(*A*0)= 0 because their support is disjoint from *A*0. On one hand, we thus have

(*s* ˜*◦ t* ˜*◦ m*)(*A*0)= ∫

= ∫

*Y*

*sy*(*A*0) d(*t* ˜*◦ m*) *sy*(*A*0) d(*t* ˜*◦ m*)+ ∫

*sy*(*A*0) d(*t* ˜*◦ m*)

*B*0 *Bc*

∫

= *sy*(*A*0) d(*t* ˜*◦ m*)

*B*0

∫

*≤*

*B*0

= 0*.*

*sy*(*A*) d(*t* ˜*◦ m*)

On the other hand, since we have *p*(*A*0)+ *p*(*Aє*)= *p*(*A*) *>* 0 and *Aє ⊂ f−*1(*Bє*), it suffices to show *p*(*f−*1(*Bє*)) = 0 to conclude *p*(*A*0) *>* 0.

By hypothesis, we have

(*s* ˜*◦ q*)(*A*)= ∫*B*0 *sy*(*A*) d*q* + ∫*Bc sy*(*A*) d*q* = 0*,*

so *q*(*Bє*) = 0 and because *f* is measure preserving, we have *p*(*f−*1(*Bє*)) = *q*(*Bє*)=0 as desired.

So (*g ◦ f, s* ˜*◦ t*) is not absolutely coherent. This completes the proof of this case.

**Lemma A.2** *Given*

*ƒ*

(*X, p*)

*ƒ*

(*Y, q*)

(*X, pj*)

*where f is a continuous function preserving the measure of both Borel probability*

*measures p and pj. If p pj, then* d*py*

d*p*

*y*

*is deﬁned q-almost everywhere and*

d*py* (*x*)= d*p* (*x*) *pj-almost everywhere.*

d*pj* d*pj*

*y*

**Proof.** For an arbitrary measurable function *h* : *X →* [0*, ∞*], by first applying the Radon-Nikodym theorem and then the disintegration theorem on the measurable function *h*  d*p* , we get

d*p′*

*h* d*p* =

∫

∫

*X* *X*

d*p h* d*pj*

d*p* =

∫

*Y*

∫*ƒ−*1(*y*)

d*p h* d*pj*

d*pj*

*y*

d*qj.*

Hence, for *q*-almost every *y*, we must have d*py* (*x*)=  d*p* (*x*) *pj*-almost everywhere.*2*

d*p′* d*p′*

*y*